SIMULADO DE PROVA

APRENDIZADO DE MÁQUINA SUPERVISIONADO

Conceitos Gerais

- O que é o aprendizado de máquina supervisionado?
- O que é aprendizado não-supervisionado?
- O que são atributos e classes?
- De um exemplo de uma instância (amostra) de um problema de classificação qualquer. Por exemplo, como você classificaria carros? E cães e gatos?
- O que significa a anotação dos dados?
- O que significa representativade em termos de características ?
- Dado o dataset abaixo, determine o que são atributos e o que são classes (target):

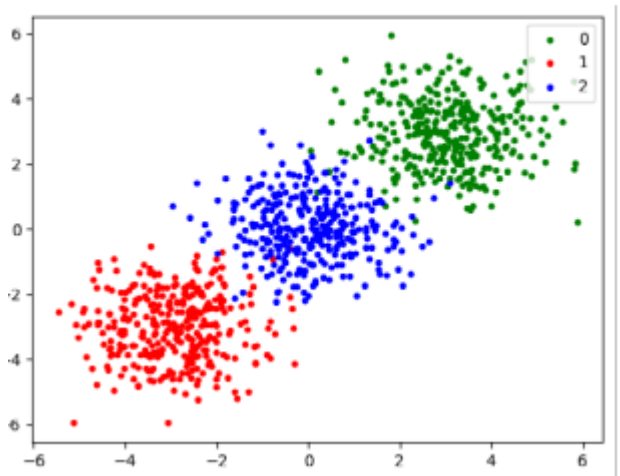
Feature 1	Feature 2	Feature 3	Cor	Tipo de Veículo
1500	110	30	Vermelho	Moto
2500	150	50	Azul	Carro
3500	220	70	Verde	Caminhão
1200	80	20	Preto	Moto
2800	160	60	Branco	Carro
4200	240	90	Vermelho	Caminhão
800	60	15	Preto	Moto
2000	120	40	Azul	Carro
3000	200	80	Verde	Caminhão

- Quais etapas possui um fluxo (pipeline) de aprendizado de máquina?
- O que é classificação binária e multi-classes?
- Como abordar classificação multi-classes a partir de modelos binários?
- Diferencie a técnica One vs One e One vs All

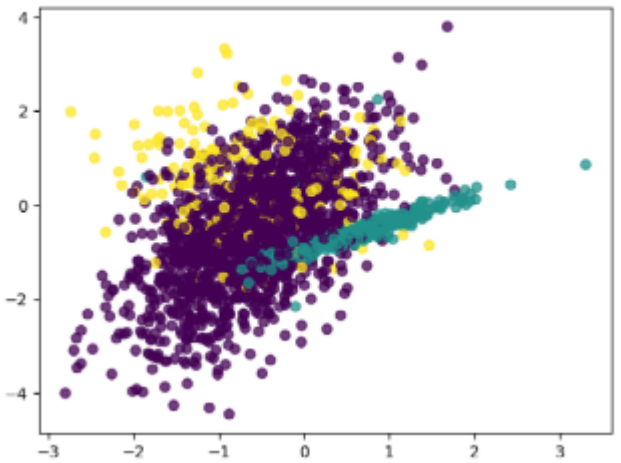
Análise Exploratória

- O que significa a análise exploratória dos dados? O que desejamos verificar com isso
- O que são dados categóricos e dados númericos?
- Como converto um dado categórico em númerico? De um exemplo.

- Considerando o dataset de veículos acima, converta o atributo 'Cor' em número utilizando a técnica "one hot enconding".
- O que é um dataset desbalanceado? O desbalanceamento ocorre em termos de atributos ou classes?
- Que técnicas podem minimizar o impacto de dados desbalanceados? Quando utiliza-las?
- Na técnica Oversampling, o que significa interpolar as amostras? Ilustre um exemplo
- Análisando as distribuições abaixo (A e B):
    - Qual aprensenta as fronteiras decisão mais definidas? Justifique sua resposta
    - Qual apresenta desbalanceamento de classes? Justifique sua resposta
- A



B



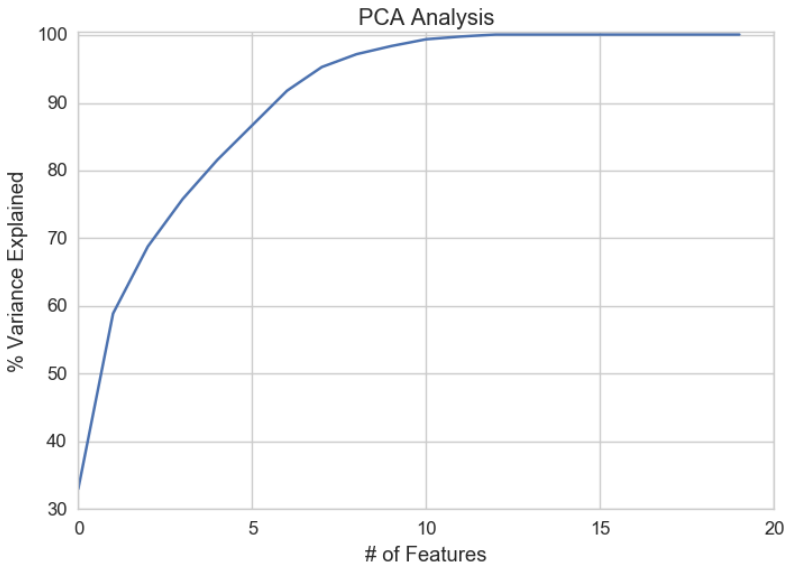
O que é a normalização de atributos? Porque isso é importante? Normalize as features abaixo por minmax():

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Feature 2	Feature 3
110	30
150	50
220	70
80	20
160	60
240	90
60	15
120	40
200	80

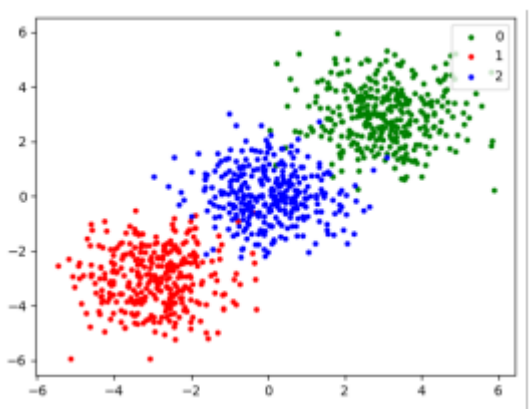
- O que é a redução de atributos e quando devemos aplicá-la?

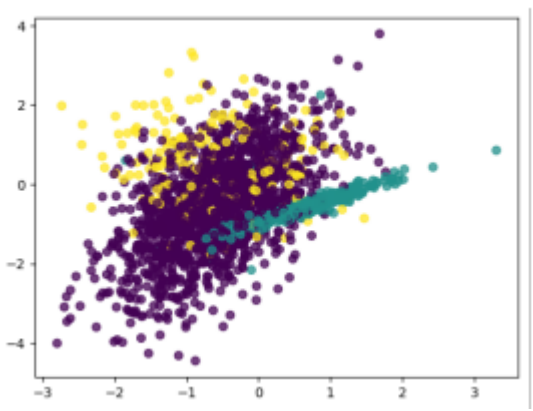
- O que significa a correlação de atributos? Dê exemplos.
- O que faz o algoritmo PCA?
- Como interpretar o grafico da variância (PCA) abaixo?



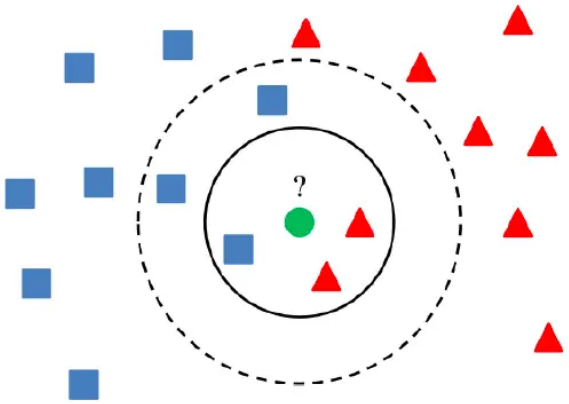
Algoritmo KNN

- Descreva em poucas linhas o algoritmo KNN. Se preferir, faça um desenho auxiliar e explique.
- O KNN funciona somente para 2 classes (binário) ?
- Qual a desvantagem do KNN em datasets grandes ? Por exemplo, com 100 mil amostras e 2000 atributos?
- O KNN reduz o espaço de características? E o espaço de busca? Justifique.
- O que é o parametro K, do KNN? Como ele impacta na classificação? Justifique.
- Analisando as seguintes distribuições, em qual o algoritmo KNN deve performar melhor? Porque?
    - Qual o impacto de um K maior e menor em cada uma das distribuições?





- Qual a classe da amostra de teste para K=3 e K=5, abaixo?



Considere o seguinte dataset e a amostra de teste abaixo:

Exemplo	Característica 1	Característica 2	Característica 3	Classe
1	2	3	1	A
2	1	2	0	A
3	3	0	2	B
4	0	1	3	B
5	2	2	3	A

Amostra de Teste	Característica 1	Característica 2	Característica 3
1	2	1	2

Utilizando a distância Euclidiana abaixo, qual o resultado da amostra para K=1 e K=3?

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

Algoritmo Naive Bayes

- Explique de maneira sucinta como funciona o Naive Bayes?
- Explique sucintamente o teorema de bayes. Cite exemplos.
- O que é uma probabilidade a posteriori e a priori? Como isso é aplicado no Naive Bayes?
- Dado dataset de frutas abaixo:

ID	Cor	Tamanho	Tipo de fruta	Classe
1	Verde	Pequeno	Limão	Ácido
2	Amarelo	Médio	Abacaxi	Doce
3	Laranja	Médio	Laranja	Ácido
4	Amarelo	Pequeno	Banana	Doce
5	Verde	Grande	Melancia	Doce
6	Vermelho	Pequeno	Morango	Doce
7	Amarelo	Médio	Pêra	Doce
8	Laranja	Grande	Tangerina	Ácido
9	Amarelo	Pequeno	Limão	Ácido
10	Verde	Médio	Maçã	Doce
11	Verde	Pequeno	Limão	Ácido
12	Amarelo	Médio	Abacaxi	Doce
13	Laranja	Médio	Laranja	Ácido
14	Amarelo	Pequeno	Banana	Doce
15	Verde	Grande	Melancia	Doce
16	Vermelho	Pequeno	Morango	Doce
17	Amarelo	Médio	Pêra	Doce
18	Laranja	Grande	Tangerina	Ácido
19	Amarelo	Pequeno	Limão	Ácido
20	Verde	Médio	Maçã	Doce

Aplique o algoritmo Naive Bayes para determinar a probalidade e classes das amostras abaixo:

ID	Cor	Tamanho	Tipo de fruta
1	Verde	Pequeno	Limão
2	Vermelho	Médio	Maçã

- Como aplicar o modelo Naive Bayes datasets com atributos númericos, tais como como peso, altura, salario, etc? De exemplos.

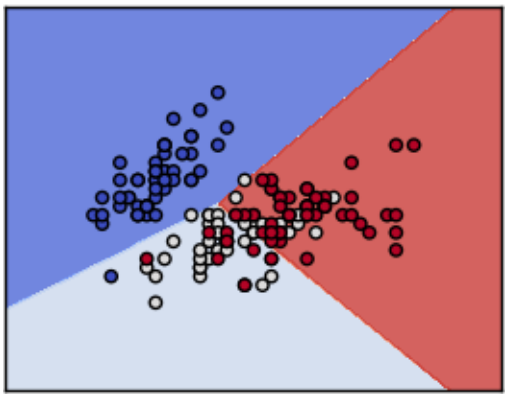
Algoritmo Decision Tree

- Explique com suas palavras o algoritmo de árvore de decisão. Ilustre um exemplo
- O que é a entropia, probabilidade e ganho de informação?
- O que o ganho de informação representa para este algoritmo?
- Considerando o dataset de frutas acima, calcule a entropia e o ganho de informação para a característica 'Tamanho'.

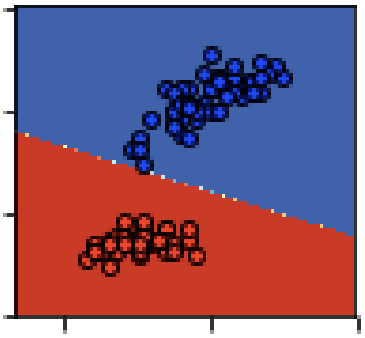
Análise Crítica

- Analisando as fronteiras de decisão, o que se pode infererir quanto a generalização do modelo? Quais classes devem sofrer perdas de acurácia e porque?

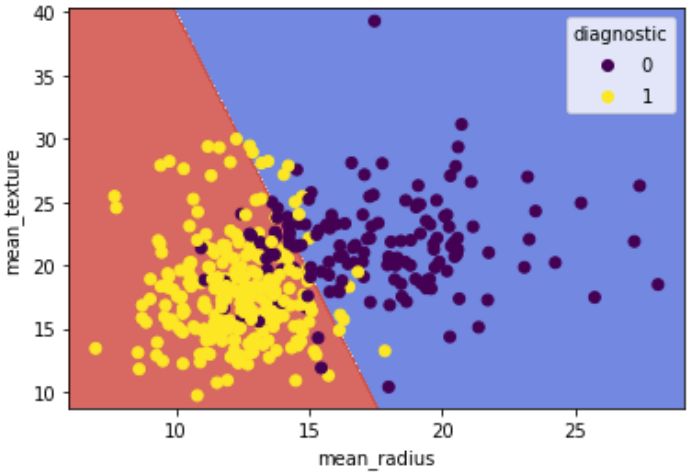
Modelo A



Modelo B



Modelo C



- O que determinas as métricas de acurácia e recall?
- Porque a acurácia geral não é uma boa métrica? Dê um exemplo.
- Calcule a acurácia a partir da matriz de confusão abaixo:

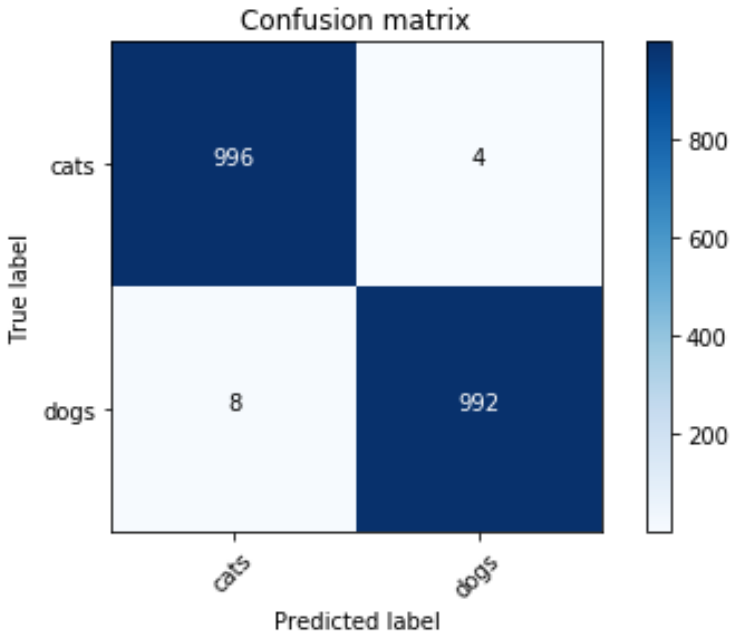
		Prediction	
		Cat	Dog
Actual	Cat	15	35
	Dog	40	10

- Dados as matrizes de confusão abaixo, análise os casos individualmente quanto:
  - Qual a acurácia global ?

O modelo está bem ajustado ou existe overfitting?

O dataset pode ser considerado balanceado?

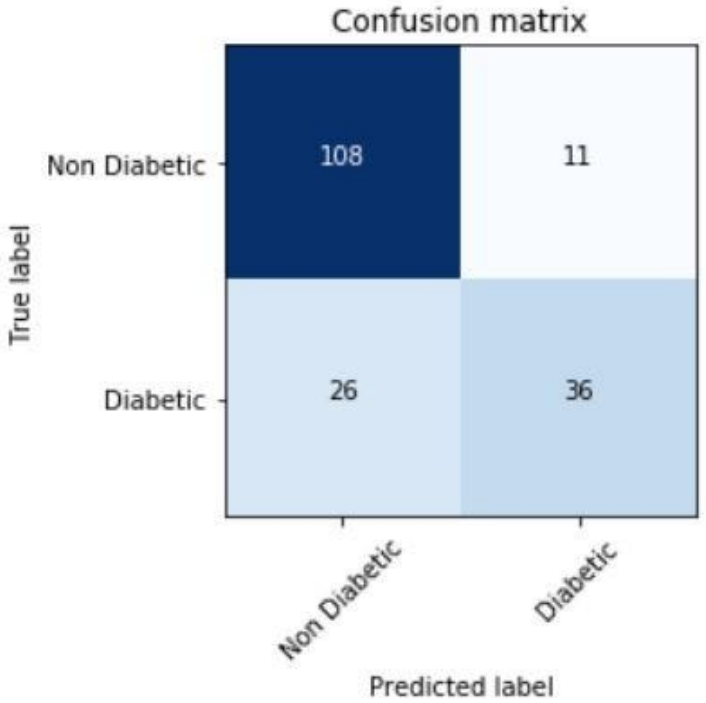
Caso A:



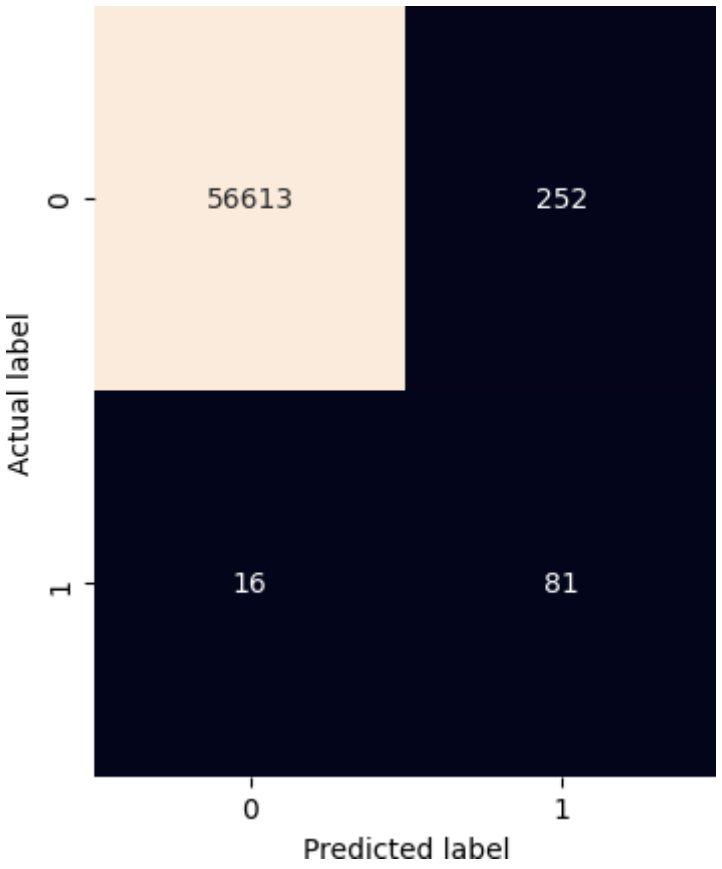
Caso B:

n=192	**Predicted: 0**	**Predicted: 1**
Actual: 0	118	12
Actual: 1	47	15

Caso C:


Confusion matrix
Non Diabetic
Diabetic
True label
108
11
26
36
Non Diabetic
Diabetic
Predicted label

Caso D:


0
1
Actual label
56613
252
16
81
0
1
Predicted label